Reporte final

Eduardo Joel Cortez Valente A01746664

Ana Martínez Barbosa A01382889

José María Ibarra A01706970

Jorge Isidro Blanco Martínez A01745907

Maximiliano Benítez Ahumada A01752791

TC3006C. Inteligencia artificial avanzada para la ciencia de datos

15/09/2023

Contexto

La analítica de datos en la venta de bienes raíces puede proporcionar información valiosa tanto para los vendedores como para los compradores , ya que de esta manera se puede tener precios acordes a la propiedad, lo que hace que los compradores paguen lo justo tomando en cuenta las características de la casa y los vendedores pueden vender las propiedades en menor tiempo.

Las agencias inmobiliarias tendrán una clara ventaja competitiva frente a aquellas que no lo utilicen porque podrán hacer ventas más precisas proporcionando datos duros a los clientes, lo que los puede dejar más satisfechos y generar más ventas. Por lo tanto, La analítica de datos no solo aumenta la eficiencia y la precisión en la toma de decisiones, sino que también mejora la experiencia del cliente y proporciona información relevante sobre el mercado, lo cual puede ser crucial al momento de realizar una venta.

I. Descripción Base de Datos

Para la obtención de un modelo que permitiese generar predicciones acertadas, se utilizó una base de datos llamado ***Ames Housing Dataset***, obtenido de la plataforma Kaggle. Dicho dataset, muestra un conjunto de 81 columnas, de las cuales 80 son características pertenecientes a 1460 propiedades ubicadas en Ammes, Estados Unidos.

Dentro del set de datos se encontraban variables categóricas (nominales y ordinales) y numéricas (discretas y continuas), a continuación se presenta una tabla con la clasificación y significado de cada una.

Variable	Descripción	Tipo de Dato	Tipo de Variable
Id	Id unico del dato	Entero	Numérico discreto
MSSubClass	Tipo de vivienda	Entero	Categórico nominal
MSZoning	Clasificacion de la zona	Objeto	Categórica nominal
LotFrontage	Distancia en pies que conecta a la propiedad con la calle	Flotante	Numérico continuo
LotArea	Área del lote en pies cuadrados	Entero	Numérico discreto
Street	Calle donde está ubicada la vivienda	Objeto	Categórica nominal
Alley	Tipo de callejón de entrada	Objeto	Categórica nominal
LotShape	Forma general de la propiedad	Objeto	Categórica nominal
LandContour	llanura de la propiedad	Objeto	Categórica nominal
Utilities	Utilidades disponibles	Objeto	Categórica nominal
LotConfig	Configuración de lote	Objeto	Categórica nominal
LandSlope	Pendiente	str	Categórica nominal
Neighborhood	Vecindario	Objeto	Categórica nominal
Condition1	Proximidad a calle principal 1	Objeto	Categórica nominal
Condition2	Proximidad a calle principal 2	Objeto	Categórica nominal
BldgType	Tipo de construcción	Objeto	Categórica nominal
HouseStyle	Tipo de vivienda	object	Categórica Nominal
OverallQual	Calidad general del material y del acabado	int64	Numérica Ordinal
OverallCond	Calificación de condición general	int64	Numérica Ordinal
YearBuilt	Fecha de construcción original	int64	Numérica Discreta
YearRemodAdd	Año de remodelación	int64	Numérica Discreta
RoofStyle	Tipo de techo	object	Categórica Nominal
RoofMatl	Material del techo	object	Categórica Nominal
Exterior1st	Revestimiento (material) del exterior	object	Categórica Nominal
Exterior2nd	Revestimiento (material) del exterior, si es más de un material	object	Categórica Nominal
MasVnrType	Tipo de chapa de mampostería	object	Categórica Nominal
MasVnrArea	Área de revestimiento de mampostería en pies cuadrados	float64	Numérica Continua
ExterQual	Calidad de materiales exteriores	object	Categórica Ordinal
ExterCond	Estado actual del material en el exterior	object	Cat
Foundation	Tipo de cimiento	object	Categórica Nominal
BsmtQual	Altura del sótano	object	Categórica Ordinal
BsmtCond	Estado general del sótano	object	Categórica Ordinal
BsmtExposure	Muros de sótano a nivel de salida o jardín	object	Categórica Ordinal
Heating	Calefacción	Objeto	Categórica nominal
HeatingQC	Calidad de Calefacción	Objeto	Categórica ordinal
CentralAir	Aire Acondicionado Central	Objeto	Categórica nominal
Electrical	Sistema Eléctrico	Objeto	Categórica nominal
1stFlrSF	Área del Primer Piso	Entero	Numérico discreto
2ndFlrSF	Área del Segundo Piso	Entero	Numérico discreto
LowQualFinSF	Área de Acabado de Baja Calidad	Entero	Numérico discreto
GrLivArea	Área de Espacio Habitacional (sobre nivel del suelo)	Entero	Numérico discreto
BsmtFullBath	Baños Completos en Sótano	Entero	Numérico discreto
BsmtHalfBath	Medios Baños en Sótano	Entero	Numérico discreto
FullBath	Baños Completos	Entero	Numérico discreto
HalfBath	Medios baños	int	Numérico discreto
BedroomAbvGr	número de cuartos	int	Numérico discreto
KitchenAbvGr	número cocinas	int	Numérico discreto
KitchenQual	Calidad cocina	object	Categórico ordinal
TotRmsAbvGrd	Total cuartos	int	Numérico discreto
Functional	Home functionality	object	Categorica dicreta
Fireplaces	Cantidad fireplaces	int	Numérico discreto
FireplaceQu	calidad fireplace	object	Ctegórico nominal
GarageType	Ubicación del garage	object	Categórico nominal
GarageYrBlt	Años que lleva construido	int	Numérico continuo
GarageFinish	Acabados interior	object	Categórica nominal
GarageCars	Tamaño garage por cantidad coches	int	Numérico discreto
GarageFinish	Acabados interior	object	Cat
GarageCars	Tamaño garage por cantidad coches	int	Numérico discreto
WoodDeckSF	Área de la cubierta	int64	Numérico discreto
OpenPorchSF	Área del porche	int64	Numérico discreto
EnclosedPorch	Área de porche cerrado	int64	Numérico discreto
3SsnPorch	Área de porche de tres estaciones	int64	Numérico discreto
ScreenPorch	Área de porche con malla	int64	Numérico discreto
PoolArea	Área de alberca	int64	Numérico discreto
PoolQC	Calidad de alberca	object	Categórico ordinal
Fence	Calidad de barda	object	Categórico ordinal
MiscFeature	Características misceláneas	object	Categórico nominal
MiscVal	Valor de las características misceláneas	int64	Numérico discreto
MoSold	Mes de venta	int64	Categórico nominal
YrSold	Año de venta	int64	Categórico nominal
SaleType	Tipo de venta	object	Categórico nominal
SaleCondition	Condición de venta	object	Categórico nominal
SalePrice	Precio de venta	int64	Numérico discreto

Figura 1. Fragmento de código de descripción y clasificación de cada variable.

II. Preprocesamiento

La construcción de cualquier modelo requiere de un proceso de limpieza y análisis previo; preprocesamiento. Este proceso depende fundamentalmente de la naturaleza de los datos con los que se esté trabajando, por lo que cada proceso es diferente y requiere de una explicación y justificación de las medidas tomadas en cada paso.

El preprocesamiento de nuestros datos comienza con un análisis de las características principales de estos. En concreto, poder diferencias entre datos numéricos y datos categóricos, los cuales a su vez estarán agrupados en sus respectivas sub-clasificaciones. Una vez agrupados, se puede realizar diferentes observaciones a estos; de los datos numéricos obtuvimos las medidas de tendencia, mientras que para los categóricos obtuvimos las modas. Para ambos tipos de datos calculamos las distribuciones de frecuencia, lo que cuál es fundamental para tomar decisiones en invitaciones, y para entender la cardinalidad de los mismos.

Posteriormente, se efectuó un trabajo con los datos nulos. Para ese punto es necesario entender que no porque haya apariciones de datos nulos se les puede dar tratamiento igual a las columnas, y es necesario tomar en cuenta el contexto de los datos (lo que miden) para definir qué tratamiento recibirán. Primero, prescindimos de las columnas que presentaban más de 500 datos nulos, lo que sería

más de un tercio de las apariciones. Luego, tiramos las variables que estaban altamente correlacionadas, donde nuestro límite de decisión fue una correlación mayor a 0.80. Este valor es arbitrario, pero es un límite que en convención indica correlaciones considerables. Después, tiramos variables con baja cardinalidad, esto es, variables que tenían muchos de sus registros acomodados en una misma medida. También tiramos variables que podrían propiciar *data leaking*, como lo pueden ser variables del detalle de la compra, las cuales se registran después del precio, el cual es nuestro *target*. Las variables borradas en este poso fueron: *'Street', 'LandContour', 'Utilities', 'LandSlope', 'Condition1', 'Condition2', 'BsmtCond', 'RoofMatl', 'Heating', 'CentralAir', 'Electrical', 'Functional', 'GarageQual', 'BsmtFinType2', 'GarageCond', 'PavedDrive', 'SaleType', 'SaleCondition', 'BldgType', 'ExterCond', 'BsmtFullBath'*.

En este punto, tomamos medidas de ingeniería de datos, donde extrajimos información en la columna de *YearRemodAdd* para establecer si una vivienda había sido remodelada sin tener información redundante (año de construcción). En este punto también codificamos las variables ordinales. Esto es importante porque el siguiente paso corresponde a la imputación de valores nulos restantes, donde se usó la moda. Sin embargo, es necesario mencionar que no se imputaron modas en todos los valores nulos, pues muchos de estos correspondían a un valor de escala en el contexto de los valores ordinales y fue necesario codificarlos según ese orden.

El último paso considerable de este preprocesamiento fue una técnica llamada *mean encoder*. La cual agrupa las apariciones únicas por columnas y calcula la media de esas apariciones según su valor de la columna objetivo, en este caso *SalePrice*. Esta columna extrae información del *target*, por lo que debe usarse con cautela. Sin embargo, puede ser apropiada en casos de regresión, donde queremos capturar la relación directa entre la variable y un valor continuo. Finalmente, antes de usar los datos para el entrenamiento, se estandarizaron los datos con la técnica *Min-Max*, de manera que el modelo no se viera sesgado por los datos que naturalmente tiene escalas más grandes.

III. Generación del Modelo

Una vez limpios los datos, se implementó el modelo de machine learning que permitiría obtener las predicciones más acertadas. Para el entrenamiento local del modelo y la evaluación con *k-fold cross validation* se dividió el conjunto de datos en 70 % de entrenamiento y 30 % de prueba.

Se decidió utilizar un perceptrón multicapa (MLP Regressor) el cual es una red neuronal artificial que utiliza *backpropagation* para ajustar los pesos entre las neuronas para mejorar la accuracy del modelo.

Dicho modelo es útil tanto para regresiones cuya salida es continua como discreta. Así mismo, su arquitectura consiste en una capa de entrada, capas ocultas y finalmente una capa de salida. En nuestro caso, decidimos implementar 3 capas ocultas, todas de 34 neuronas. Finalmente, como se quiere predecir un solo resultado, la capa de salida es de una sola neurona.

Para la función de activación, se utilizó la función *ReLu*, la cual maneja la no linealidad. Además, se implementó '*LBFGS*' como *solver* para ajustar los pesos. Dicha elección proviene de que es un *solver* que se desenvuelve bastante bien antes muchos parámetros (nosotros tenemos 80 más las columnas de *dummies*) y es usado en regresiones.

Por otra parte, para evitar overfitting en nuestro modelo se utilizó como técnica de regularización '*L2*', con un *alpha* de 1.3 obtenido a través de diversas pruebas con diversos valores.

```
('mlp', MLPRegressor(hidden_layer_sizes=(34, 34, 34),
                     alpha=1.3,
                     activation='relu', solver='lbfgs',
                     random_state=10))
```

Figura 2. Fragmento de código de MLP Regressor

Finalmente, para evaluar el rendimiento de nuestro modelo se utiliza el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE) y el R^2 ajustado.

```
R2 score: 0.8722619496266457
MSE: 749002488.5252677
RMSE: 27367.90983113741
```

Figura 3. Resultados del modelo.

Se puede apreciar que el modelo explica con 86% de certeza los datos, por lo cual tiene una confiabilidad bastante grande. Así mismo, su error, comparado con los precios de las casas, es bastante bueno.

Por último, para poder implementar el modelo en nuestro objetivo final, el cual era una aplicación, se debieron de ajustar la cantidad de parámetros a considerar dentro del modelo.

Para el usuario no sería útil ni atractivo el encontrar un cuestionario con las 34 variables finales del modelo. Así mismo, no todas las variables eran fáciles de obtener la información para el usuario, por ejemplo la cantidad de pies que comprendía el recubrimiento de la casa. Por otro lado, había otras variables que eran subjetivas o difíciles de responder, por ejemplo si se esperaba que el usuario clasifique la calidad de su casa, de sus acabados o de un área específica.

Por ende, se optó por, a partir de los resultados de los coeficientes arrojados por el modelo, reducir

aún más su cantidad.

Se optó por dejar 15 variables en el cuestionario y, para no afectar el modelo, las 19 variables más del modelo se mantuvieron con valores estáticos.

Dichas 15 variables debían cumplir con ser de las más significativas para el modelo y de poder ser amigables con el usuario.

Descripción del producto final

I. ¿Cómo resolvemos la problemática?

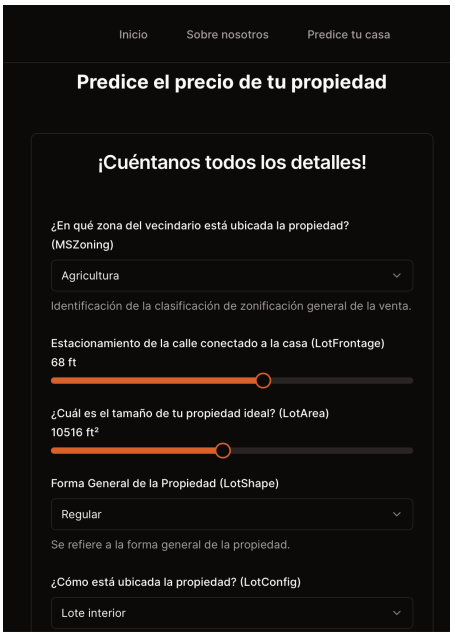
Describiendo a detalle el modelo previamente presentado, podemos afirmar que este resuelve con creces la problemática inicialmente planteada. En este sentido, ¿cómo transformamos dicha solución técnica en un producto usable y económicamente viable? Así es como concebimos NeighborHub.

Comenzamos definiendo NeighborHub como una aplicación web adaptable, puesto que esto nos concedería presencia en una amplia gama de dispositivos; por ende, una gran cantidad de usuarios.

Una vez establecido el tipo de producto a desarrollar, se eligió Python Flask y Next.js como tecnologías para Backend y Frontend respectivamente.

Así pues, dichas herramientas nos permitieron generar una aplicación profesional, amigable e intuitiva, cuyo propósito principal es permitir al cliente conocer el precio estimado de una propiedad a comprar/vender de la manera más precisa posible.

Conseguimos cumplir el objetivo principal de la plataforma mediante un formulario que recibe las 15 variables establecidas en la descripción del modelo utilizado.



II. Modelo de negocios

Como primer paso de nuestro escalamiento continuo, la plataforma será gratuita para todo tipo de usuarios que deseen predecir el precio de una propiedad a vender o comprar. En este sentido, se tiene prevista la publicidad como medio para generar fondos y así continuar con el desarrollo de la aplicación. Así pues, se estará trabajando exhaustivamente en la creación y optimización de modelos de ML más eficientes y exactos para implementarse.

En este sentido, una vez se cuente con más modelos lo suficientemente robustos, se le ofrecerá al cliente la posibilidad de aportar una pequeña

contribución monetaria a cambio de requests que ofrecerán mucha más precisión.

III. Normatividad de la solución implementada

NeighborHub se compromete a cumplir con todas las leyes y regulaciones en la industria inmobiliaria y tecnológica. Esto incluye:

1. Protección de Datos Personales: Garantizan la privacidad y seguridad de los datos de los usuarios, cumpliendo con leyes como el RGPD. Proporcionan controles para la gestión transparente de datos personales.

2. Transparencia en la Recopilación de Datos: Informan a los usuarios sobre qué datos se recopilan y cómo se utilizan mediante políticas de privacidad claras.

3. No Discriminación: La plataforma es inclusiva y no discrimina a los usuarios por características personales.

4. Exactitud de los resultados: Se esfuerzan por proporcionar estimaciones de precios precisas, minimizando sesgos y utilizando datos confiables.

5. Publicidad Responsable: Compromiso de mostrar anuncios éticos y no invasivos que no afecten la experiencia del usuario.

Cumplimiento de Leyes, Normas y Principios Éticos:

NeighborHub se compromete a cumplir con todas las leyes y regulaciones aplicables en la industria inmobiliaria y en el ámbito de la tecnología de la información. Para garantizar la ética y la transparencia en nuestro servicio, hemos establecido las siguientes medidas:

Normatividad Correspondiente:

A lo largo del proceso de desarrollo, así como el tiempo de operación de la plataforma, hacemos énfasis en la normatividad específica relacionada con la industria inmobiliaria y la tecnología en cada región donde operamos. Esto incluye la correspondiente investigación de regulaciones inmobiliarias locales, estándares de seguridad de datos y cualquier otro requisito legal relevante. Además, proporcionamos información sobre nuestras prácticas que cumplen con estas regulaciones para garantizar la transparencia y la confianza de nuestros usuarios. Estamos comprometidos a mantenernos actualizados y ajustar nuestras operaciones según sea necesario para cumplir con los cambios en la normativa y los principios éticos de la industria y/o región donde operamos.

Conclusiones.

A partir de lo generado en el transcurso de este proyecto, nos es posible llegar a múltiples conclusiones. Tanto relacionadas con el análisis de datos y el uso de herramientas de inteligencia artificial para su estudio, así como relacionadas con el producto desarrollado a partir de la generación del modelo.

Primeramente, la importancia de la analítica de datos en el mercado inmobiliario. Con lo visto y generado hasta el presente día nos fue posible ver el papel fundamental que el análisis de datos tiene en el mercado inmobiliario; al proporcionar información valiosa para vendedores y compradores. Esta información permite precios justos y eficientes, lo que beneficia a ambas partes y contribuye a una mejora en la experiencia de compra.

Así pues, las agencias inmobiliarias que adoptan la analítica de datos tienen una mayor ventaja competitiva sobre aquellas que no. Es resaltable el enfoque tomado sobre el proyecto para proporcionar datos concretos y predicciones precisas.

Nuestra elección de un perceptrón multicapa como modelo de machine learning es un enfoque efectivo para predecir valores en el mercado inmobiliario. La alta tasa de certeza que tuvo demuestra que la arquitectura de un modelo con capas ocultas y una capa sólida; junto con la función de activación ReLu y el solver LBFGS, en su conjunto, resultan útiles en este contexto en concreto.

Por último, nos parece pertinente señalar que en última instancia la analítica de datos en cualquier industria es un proceso en constante evolución. La mejora continua, tanto en la elección de datos como en la adaptación del modelo a las necesidades cambiantes de los usuarios, es esencial para mantener una ventaja competitiva en estes sector; por lo cual, creemos que el trabajo realizado hasta este punto son solo pasos iniciales para un producto en constante crecimiento y mejora.